Description

Speech recognition system, training arrangement and method of calculating iteration values for free parameters of a maximum-entropy speech model

The invention relates to a method of calculating iteration values for free parameters $\lambda_{\alpha}^{ortho(n)}$ of a maximum-entropy speech model MESM in a speech recognition system with the aid of the generalized iterative scaling training algorithm in accordance with the following formula:

$$\lambda_{\alpha}^{ortho(n+1)} = G(\lambda_{\alpha}^{ortho(n)}, m_{\alpha}^{ortho}, ...) \qquad (1)$$

where:

n : is an iteration parameter;

G : is a mathematical function;

α : is an attribute in the MESM; and

m_{α}^{ortho} : is a desired orthogonalized boundary value in the MESM for the attribute α.

The invention further relates to a computer-supported speech recognition system known in the state of the art, as well as a known computer-supported training arrangement in which the method described is implemented.

A starting point for the formation of a speech model as it is related in a computer-supported speech recognition system for recognizing entered speech is a predefined training object. The training object maps certain statistical patterns in the language of a future user of the speech recognition system into a system of mathematically formulated boundary conditions, which system generally has the following form:

$$\sum_{(h,w)} N(h) \cdot p\ (w \mid h) \cdot f_{\alpha}(h,w) = m_{\alpha} \qquad (2)$$

where:

N(h) : refers to the frequency of history h in a training corpus;

$p_{(w|h)}$: refers to probability p(w | h) with which a predefined word w follows a previous word sequence h (history);

$f_{\alpha}(h,w)$: refers to a binary attribute function for an attribute α; and α

m_{α} : refers to a desired boundary value in the system of boundary conditions.

The solution of this system of boundary conditions i.e. the training object is formed by the so-termed maximum-entropy speech model MESM which indicates a suitable solution of the system of boundary conditions in the form of a suitable definition of the probability p(w | h), which reads as follows:

$$p(w|h) = p\lambda^{(w|h)} = \frac{1}{Z_{\lambda}(h)} \cdot \exp\left(\sum_{\alpha} \lambda_{\alpha} \cdot f_{\alpha}(h,w)\right) \qquad (3)$$

where:

Zλ (h) : refers to a history-dependent standardization factor;

λα : refers to a free parameter for the attribute α;

λ : refers to the set of all parameters. For the above parameters hold their above definitions.

The binary attribute function fα(h,w) makes, for example, a binary decision whether predefined word sequences h,w contain predefined words at certain locations. An attribute α may generally refer to a single word, a word sequence, a word class (color or verbs), a sequence of word classes or more complex patterns.

Fig. 4 shows predefined attributes in a speech model by way of example. For example, the unigrams shown each represent a single word, the bigrams each represent a word sequence consisting of two words and the trigram shown represents a word sequence consisting of three words. The bigram "ORA" includes the unigram "A" and, in addition, includes a further word; therefore it is referred to as having a larger range compared to the unigram "A". Analogously, the trigram "A WHITE HOUSE" has a larger range than the unigram "HOUSE" or than the bigram "WHITE HOUSE".

The free parameters λ are adapted so that equation 3 represents a solution for the system of boundary conditions according to equation 2. This adaptation is normally made with the aid of known training algorithms. An example for such a training algorithm is the so-termed generalized iterative scaling GIS algorithm as it is described, for example, in J.N. Darroch and D. Ratcliff, "Generalized iterative scaling for log linear models", Annals Math. Stat., 43(5):1470-1480, 1972.

This GIS algorithm provides an iterative calculation of the free parameters λ. Traditionally, this calculation is made very slowly, however. For expediting this calculation, there is proposed in the state of the art to substitute orthogonalized attribute functions f_α^{ortho} (h,w) for the attribute functions f_α (h,w) in the system of boundary conditions in accordance with equation (2); see for this purpose R. Rosenfeld "A maximum-entropy approach to adaptive statistical language modeling"; Computer Speech and Language, 10:187-228, 1996. Because of the substitution of the attribute functions on the left in equation 2, however, also the boundary values m_α on the right are changed. This changes the original system of boundary conditions i.e. the original training object in the customary sets approaches for estimating the boundary values; for this purpose see Rosenfeld at other locations, page 205, first sentence of the last-but-one paragraph.

In this respect it can be established as a disadvantage of the state of the art that when the calculation of the GIS algorithm is accelerated, the free parameters λ are trained to a changed training object. The parameters λ calculated in this manner are the cause for an inadequate adaptation of the speech model to the original training object when the parameter λ is used in equation 3.

Starting from this state of the art it is an object of the invention to further develop a known computer-supported speech recognition system, a computer-supported training system and a known method of iteratively calculating free parameters $\lambda_\alpha^{ortho(n)}$ of a maximum-entropy speech model in the speech recognition system, so that they make a fast calculation possible of the free parameters λ without a change of the original training object.

This object is achieved as claimed in patent claim 1 in that with the known above-described method of calculating the free parameters λ according to the GIS algorithm, any desired orthogonalized boundary value m_α^{ortho} is calculated by linearly combining the associated desired boundary value m_α with desired boundary values m_β of attributes β that

have a larger range than the attribute α. Here m_α and m_β are desired boundary values of the original training object.

The use of the boundary values m_α^{ortho} calculated in this manner makes it possible in an advantageous manner to make an improved approximation of the free parameters λ and thus an improvement of the speech model with a view to the original training object. This qualitative improvement is possible while a high convergence speed continues to realize for the free parameters λ during the iterative calculation with the aid of the GIS algorithm.

The use of the desired orthogonalized boundary values m_α^{ortho} calculated according to the invention is recommended for several variants of the GIS training algorithm as they are described in the dependent claims 12 and 13.

The object of the invention is furthermore achieved by a speech recognition system based on the maximum-entropy speech model MESM as claimed in claim 14 and a training system for training the MESM as claimed in claim 15.

By implementing the method according to the invention in the training system, compared to the state of the art the MESM in the speech recognition system is adapted more effectively to the individual language peculiarities of a certain user of the speech recognition system; the quote with which the speech recognition system then correctly recognizes the semantic content in the user's speech is improved considerably.

Otherwise the advantages of this speech recognition system and of the training system correspond to the advantages discussed above for the method.

The following Figures are added to the description of the invention, in which

Figs. 1a and 1b describe a method according to the invention of calculating a desired orthogonalized boundary value m_α^{ortho} ;

Figs. 2a and 2b describe a method according to the invention of calculating an orthogonalized attribute function f_α^{ortho} ;

Fig. 3 describes a block diagram of a speech recognition system according to the invention;

Fig. 4 describes an attribute tree.

In the following first a detailed description is given of an example of embodiment of the invention while reference is made to Figs. 1a and 1b.

Figs. 1 and 1b illustrate a method according to the invention of calculating an improved desired orthogonalized boundary value m_{α}^{ortho} for an attribute $\alpha = \beta 0$ in a speech model. In a first step of the method all the attributes βi with $i = 1 \ldots g$ that have a so-termed larger range than the predefined attribute $\alpha = \beta 0$ i.e. which include this at a predefined position are determined in accordance with this method. Subsequently, in a method step S2 a desired boundary value $m\beta i$ of the original training object is calculated for all the attributes βi with $i = 0 \ldots g$, thus also for the attribute $\alpha = \beta 0$.

For the calculation of such a desired boundary value $m\beta i$, several methods are known in the state of the art.

According to a first method the calculation is made in that first a frequency $N(\beta i)$ is determined with which the associated binary attribute function $f\beta i$ yields the value 1 when a training corpus of the speech model is used and that, subsequently, the thus determined frequency value $N(\beta i)$ is smoothed.

According to a second, alternative method, the calculation is performed by reducing the quantities of attributes in the speech model until the boundary conditions no longer demonstrate conflicts. This sort of reduction in the quantity of attributes must be very extensive in practical situations, since otherwise the generated speech model will no longer represent a solution to the original training object. According to a third method, the calculation is made by using a so-called induced speech model as it is described in J. Peters and D. Klakow, "Compact Maximum Entropy Language Models", Proc. ASRU, Keystone, Colorado, 1999.

In a method step S3 all the attributes βi are subsequently sorted according to their range where an attribute βi that has the largest range is assigned the index $i = g$. It may then certainly happen that individual classes of ranges thus, for example, the class of bigrams or the class of trigrams are assigned a plurality of attributes βi. In these cases a plurality of attributes βi having different, but successive indices i are assigned to one and the same class of ranges i.e. these attributes then always have the same RW and belong to the same class of ranges.

For the method to be carried out, in which in the successive steps the individual attributes βi are evaluated one after the other, it is important for the attributes to be processed according to decreasing (or constant) range. In the first run of the method a start is therefore made with an attribute βi which is assigned to the highest class of ranges; preferably i is set equal to g (see method step S4 and S5 in Fig. 1a).

In a subsequent method step S6 a check is then made whether larger-range attributes βk occur with $i < k \leq g$ for the currently selected attribute βi, which include the attribute βi. With the first run the attribute βi with i = g automatically belongs to the class that has the largest range, as observed above, and therefore the query in the method step S6 is to be answered in the negative for this attribute βi. In this case the method jumps to method step S8 where a parameter X is set to zero. Then a calculation is made of an improved desired orthogonalized boundary value $m_{\beta i}^{ortho}$ for the attribute βi (with a first run with i = g) in accordance with method step S9. As can be seen there, this boundary value for the attribute β i is set equal to the desired boundary value mβi calculated in step S2, if the parameter X = 0 (this is the case, for example, during the first run).

The method steps S5 to S11 are then successively repeated for all the attributes βi-1 with i-1 = g-1 ... 0. In the method step S10 the index i is re-initialized, which is necessary, and in method step S11 a query is made whether all the attributes βi with i = 0 ... g have been processed.

For all attributes βi for which there are attributes βk with $i < k \leq g$ that have a larger range, the query in method step S6 must be answered with "Yes". The parameter X is then not set to zero but is instead calculated according to method step S7 by totaling the corresponding improved desired orthogonalized boundary values $m_{\beta k}^{ortho}$ each calculated in previous run-throughs in method step S9 for the respective attributes βk that have a larger range.

Once it has been determined in method step S11 that the desired orthogonalized boundary value $m_{\beta 0}^{ortho}$ has been calculated in method step S9, this is then output in method step S12 as m_{α}^{ortho}. The method according to the invention extensively described just now for the calculation of the improved desired orthogonalized boundary value m_{α}^{ortho} may be shortened to the following formula:

$$m_{\alpha}^{ortho} = m\alpha - \sum_{(*)} m_{\beta}^{ortho} \quad . \qquad (4)$$

The sum (*) includes all attributes β that have a larger range and contain the predefined attribute α. For calculating the boundary value m_{β}^{ortho} said formula can be used in

an almost recursive manner for each attribute β again and again until the sum term disappears for certain attributes, that is, for those with the largest range, because there are no more attributes that have a larger range for them. The desired orthogonalized boundary values for the attributes βk that have the largest range then correspond to the respective originally desired boundary values mβk.

The implementation of the method according to the invention and as shown in Fig. 1a and 1b will be further explained hereinafter while use is made of the following training corpus of a speech model used by way of example. The training corpus reads:

"THAT	WAS	A	RED
	OR	A	GREEN HOUSE
	OR	A	BLUE HOUSE
THIS	IS	A	WHITE HOUSE AND
THAT	IS	THE	WHITE HOUSE"

The training corpus consists of N = 23 individual words. It is assumed that in the speech model the desired unigram, bigram and trigram attributes are predefined according to Fig. 4.

Then, by using the normal attribute function fα for the training corpus it may be established that the unigrams, bigrams and trigrams according to Fig. 4 occur in the training corpus with the following frequencies:

Unigrams:

A	4
HOUSE	4
IS	2
OR	2
THAT	2
WHITE	2

Bigrams:

A	WHITE	1
OR	A	2
WHITE	HOUSE	2

Trigrams:

A WHITE HOUSE 1

In the example shown here the improved desired orthogonalized boundary value m_{α}^{ortho} is to be calculated for the attribute α = "HOUSE". For this purpose first according to method step S1 in Fig. 1a all attributes that have a larger range are to be determined for the attribute α. They are according to Fig. 4 the bigram "WHITE HOUSE" and the trigram "A WHITE HOUSE". According to method step S2 the normal desired boundary values are to be calculated for these attributes that have a larger range but also for the attribute α, for example, in that the respective frequencies established above are smoothed. This smoothing is effected here, for example, by subtracting the value 0,1. Thus the following normal desired boundary values are the result:

mα	: "HOUSE"	= 4 - 0,1 =	3,9
mβ1	: "WHITE HOUSE"	= 2 - 0,1 =	1,9
mβ2	: "A WHITE HOUSE"	= 1 - 0,1 =	0,9.

The attributes α, β_1, β_2 are now sorted according to their range and – starting with the widest ranging attribute – the respective improved desired orthogonalized boundary values are calculated according to formula (6) or according to method step S7-S9 in Figs. 1a and 1b:

$$m_{\beta 2}^{ortho} = m_{\beta 2} = 0.9 \qquad (5)$$

$$m_{\beta 1}^{ortho} = m_{\beta 1} - m_{\beta 2}^{ortho} = 1{,}9 - 0{,}9 = 1 \qquad (6)$$

Finally, the improved desired orthogonalized boundary value m_{α}^{ortho} is calculated for the attribute α to:

$$m_{\alpha}^{ortho} = m_{\alpha} - m_{\beta 1}^{ortho} - m_{\beta 2}^{ortho} = 3{,}9 - 1 - 0{,}9 = 2 \qquad (7)$$

The orthogonalized boundary value m_α^{ortho} calculated according to the invention makes a sufficiently accurate calculation possible of the free parameters λ and thus of the probability according to formula (1) with a view to an original training object while the calculation velocity remains the same when used in the GIS training algorithm.

Hereinafter the use of the boundary value m_α^{ortho} calculated according to the invention will be represented for three different variants of the GIS training algorithm.

With a first variant of the GIS training algorithm the mathematical function G has the following form according to equation 1 when the orthogonalized boundary value m_α^{ortho} calculated according to the invention is used:

$$\lambda_\alpha^{ortho(n+1)} = G(\lambda_\alpha^{ortho(n)}, m_\alpha^{ortho}, ,...)$$
$$= \lambda_\alpha^{ortho(n)} + t_\alpha^{ortho} \cdot \log\left(\frac{[t_\alpha^{ortho} \cdot m_\alpha^{ortho} + b_\alpha]}{[t_\alpha^{ortho} \cdot m_\alpha^{ortho(n)} + b_\alpha]} \cdot \frac{1 - \sum_\gamma [t_\gamma^{ortho} \cdot m_\gamma^{ortho(n)} + b_\gamma]}{1 - \sum_\gamma [t_\gamma^{ortho} \cdot m_\gamma^{ortho} + b_\gamma]} \right) \quad (8)$$

where:

n	:	refers to an iteration parameter;
α	:	refers to a just considered attribute;
γ	:	refers to all the attributes in the speech model;
t_α^{ortho}, t_γ^{ortho}	:	refer to the size of the convergence step;
m_α^{ortho}, m_γ^{ortho}	:	desired orthogonalized boundary values in the MESM for the attributes α and γ;
$m_\alpha^{ortho(n)}$, $m_\gamma^{ortho(n)}$	:	refers to iterative approximate values for the desired boundary values m_α^{ortho}, m_γ^{ortho}; and
bα and bγ	:	refer to constants.

The calculation of the convergence step sizes t and of the iterative approximate values for the desired boundary values m is effected – as will be shown hereinafter – by the

use of an orthogonalized attribute function f_α^{ortho} defined according to the invention, which reads as follows:

$$f_\alpha^{ortho} = f\alpha - \sum_{(*)} f_\beta^{ortho} \qquad (9)$$

It should be observed at this point that the orthogonalized attribute function f_α^{ortho} calculated according to the invention in accordance with equation 9 corresponds as regards value to the attribute function proposed by Rosenfeld at other locations. However, their calculation according to the invention is effected totally different as can be seen in Figs. 2a and 2b. The calculation method is effected analogously to the method described in Figs. 1a and 1b for the calculation of the desired orthogonalized boundary values m_α^{ortho} where only the symbol for the boundary value m is to be replaced by the symbol for the attribute function f and the parameter X by the function F. To avoid repetitions, reference is made here to the description of Figs. 1a and 1b for explanations of the method according to Figs. 2a and 2b.

With the orthogonalized attribute function f_α^{ortho} or f_β^{ortho} thus calculated according to the invention, the size of the convergence steps t_α^{ortho} and t_γ^{ortho} is calculated in equation 8 as follows:

$$t_\alpha^{ortho} = t_\gamma^{ortho} = 1/M^{ortho} \text{ with } M^{ortho} = \max_{(h,w)}\left(\sum_\beta f_\beta^{ortho}(h,w)\right) \qquad (10)$$

where Mortho for binary attribute functions f_β^{ortho} represents the maximum number of functions which yield the value 1 for the same argument (h,w).

Furthermore, with the attribute function f_α^{ortho} defined according to the invention, the iterative approximate value $m_\alpha^{ortho(n)}$ can be calculated for the desired orthogonalized boundary value m_α^{ortho} when the following equation (2) is used:

$$m_{\alpha}^{ortho(n)} = \sum_{(h,w)} N(h) \cdot p^{(n)}(w \mid h) \cdot f_{\alpha}^{ortho}(h,w) \quad ; \qquad (11)$$

where:

N(h) : refers to the frequency of the history h in the training corpus; and

$p^{(n)}(w|h)$: refers to an iteration value for the probability p(w | h) with which a predefined word w follows a previous word sequence h (history);

Here $p^{(n)}(w|h)$ uses the parameter values $\lambda_{\alpha}^{ortho(n)}$.

The use of the improved desired orthogonalized boundary value m_{α}^{ortho} calculated according to the invention is furthermore recommended for a second variant of the GIS training algorithm. Here the attributes of the MESM are subdivided into m groups Ai and for each iteration only the parameters λ_{α}^{ortho} of the attributes α from one of the groups are changed according to the following formula:

$$\lambda_{\alpha}^{ortho(n+1)} = G(\lambda_{\alpha}^{ortho(n)}, m_{\alpha}^{ortho}, \ldots)$$
$$= \lambda_{\alpha}^{ortho(n)} + t_{\alpha}^{ortho} \cdot \log\left(\frac{m_{\alpha}^{ortho}}{m_{\alpha}^{ortho(n)}} \cdot \frac{1 - \sum_{\beta \in Ai(n)} (t_{\beta} \cdot m_{\beta}^{ortho(n)})}{1 - \sum_{\beta \in Ai(n)} (t_{\beta} \cdot m_{\beta}^{ortho})} \right) \qquad (12)$$

where:

n : represents the iteration parameter

Ai(n) : represents an attribute group Ai(n) with $1 \le i \le m$ selected in the n^{th} iteration step;

α : represents a just considered attribute from the just selected attribute group Ai(n);

β : represents all attributes from the attribute group Ai(n);

t_{α}^{ortho}, t_{β}^{ortho} : represent the size of the convergence step with $t_{\alpha}^{ortho} = t_{\beta}^{ortho} = 1/M_{i(n)}^{ortho}$ with

$$M_{i(n)}^{ortho} = \max_{(h,w)} \left(\sum_{\beta \in Ai(n)} f_{\beta}^{ortho}(h,w) \right)$$

where

$M_{i(n)}^{ortho}$ for binary functions f_{β}^{ortho} represents the maximum number of functions from the attribute group Ai(n) which yield the value 1 for the same argument (h,w);

m_{α}^{ortho}, m_{β}^{ortho} : represent the desired orthogonalized boundary values in the MESM for the attributes α and β respectively;

$m_{\alpha}^{ortho(n)}$, $m_{\beta}^{ortho(n)}$: represents iterative approximate values for the desired boundary values m_{α}^{ortho}, m_{β}^{ortho}.

The group Ai(n) of attributes α whose parameters λ_{α}^{ortho} are adapted in the current iteration step, then cyclically runs through all the m groups in accordance with i(n)=n(mod m).

The use of the desired orthogonalized boundary value m_{α}^{ortho} calculated according to the invention is further recommended for a third variant of the GIS training algorithm which distinguishes itself from the second variant only in that the attribute group Ai(n) to be used for each iteration step is not selected cyclically but according to a predefined criterion $D_i^{(n)}$.

Fig. 3 finally shows a speech recognition system 10 of the type according to this invention which is based on the so-termed maximum-entropy speech model. It includes a recognition device 12 which attempts to recognize the semantic content of supplied speech signals. The speech signals are generally supplied to the speech recognition system in the form of output signals from a microphone 20. The recognition device 12 recognizes the semantic content of the speech signals by mapping patterns in the received acoustic signal on two predefined recognition symbols such as specific words, actions or events, using the implemented maximum-entropy speech model MESM. Finally, the recognition device 12 outputs a signal which represents the semantic content recognized in the speech signal and can be used to control all kinds of equipment – for example a word-processing program or telephone.

To make the control of the equipment as error-free as possible in terms of the semantic content of speech information used as a control medium, the speech recognition system 10 must recognize the semantic content of the speech to be evaluated as correctly as possible. To do this, the speech model must be adapted as effectively as possible to the linguistic peculiarities of the speaker, i.e. the user of the speech recognition system. This

adaptation is performed by a training system 14 which can be operated either externally or integrated into the speech recognition system 10. To be more accurate, the training system 14 is used to adapt the MESM in the speech recognition system 10 to recurrent statistical patterns in the speech of a particular user.

Both the recognition device 12 and the training system 14 are normally, although not necessarily, in the form of software modules and run on a suitable computer (not shown).